For those of you who don't know me, my name is Izzy, and I'm the student rep for the equality4success committee and Elizabeth asked me if I would like to do an e4s Bite seminar and I picked the topic of biases in AI because I thought it's very interesting. I don't claim to be an expert on any of this but I've done a little bit of research for the presentation and I found some things that are quite shocking and some useful information that we might all take going forward. So I hope you'll also find this quite informative.

So to start with a definition of AI, this is artificial intelligence. And it's described as the leading technology in what's called the fourth industrial revolution. And it refers to the technological advances which cover everything from biotechnology to big data that are said to be rapidly reshaping the world as we know it. And so people in science might be directly working on this, or it might be in the periphery of our interests, but I think it should be relevant to everybody.

So beginning on researching this topic: it seems that artificial intelligence has a bias problem. And this crops up in our everyday lives. A particular example here is a gender bias problem with virtual assistants. So even just thinking about the technology that we might use in our everyday lives, such as Alexa and Siri, these are often programmed with the voices of women and they have the names of women and they often have programmed flirtatiousness which you could argue is reproducing some harmful stereotypes of the classic female secretary, which is subliminally reinforcing the role of women is secondary and submissive to men.

And another interesting point to note here is that AI systems are operating on the command of their user and they have no right to refuse, obviously, because they are programmed only to obey. And so there are some arguments that again this is dangerous in creating expectations for how women should be behaving in society.

One particular example that I came across here, is that if you suggest to Samsung's virtual assistant, Bixby, ‘let's talk dirty.’ If you have the woman's voice setting, she will say, ‘I don't want to end on Santa’s naughty list.’ Whereas if you have the male voice programmed, he'll simply say, ‘I've read that soil erosion is a real dirt problem.’ So clearly this was a real programming decision, not some kind of error. That makes us start to think about biases in technology, the choices behind them and how they're impacting society.

So, from my research, I discovered that is quite common that AI systems are biased and these generally fall into two categories along race and gender lines. And one of the most high profile cases that I found was this Amazon recruitment algorithm.

So Amazon tried to use AI to look at CVs for recruiting it’s new employees, but unfortunately this recruitment algorithm discriminated if it found the word ‘woman’ in someone’s CV and it also discriminated on the mention of women's colleges. So this they found out was because it was trained on historical data where there's been preferential recruitment of males and over representation of males in the company. And in the end, they couldn't actually fix the algorithm. And it had to be dropped, which I think, you know, highlights this is a very topical and big issue.

So thinking about gender biases, there's been a lot of research that has linked the development of these AI systems which have gender biases in them to the lack of women in the design teams. And there's an institute called AI Now Institute and they've released a report saying that there's a clear connection between the male dominated AI industry and its discriminatory systems and products.

What there's been less studies and information published on, is actually how the products themselves are incorporating stereotyped representations of gender in their design. And this is something that needs to be researched further.

A very interesting example here is in word embeddings. So this is a common AI feature where AI is able to fill the sentence ‘man is to king as women is to x’ and it fills it with the word queen. And the way this works is that words are first converted to numbers or numerical representations which are used as the input models and they are then represented as a sequence or a vector of numbers and if two words have similar meanings their embeddings will be mathematically close to one another.

And so this obviously has a lot of uses, but the issues arise for biases when AI will fill in the sentence like ‘father is to doctor as mother is to nurse’. And I'm sure everyone will be aware of the problems that these kind of associations can cause, out there in the world.

When I was researching this, this reminded me of the Harvard Project Implicit test. There is some debate in the diversity field as to the usefulness of this. However, I've found that it's quite useful to sort of have a go and see if you yourself have any unconscious biases. It's an online computer questionnaire where you click through and I personally would recommend it, however, it might be if you interested in it to check out in the literature and because their debates about how useful is. But I just thought I would flag that here.

Turning now to racial bias in AI, some of the things that I found here are really quite shocking. I wanted to highlight quite a few examples here to you, just to show you sort of the things that are happening around this issue.

So the first example that I came across was a chat bot. This is where Microsoft designed something called Tay and this is a Twitter bot which was used for conversational understanding. So the idea was that the more you chat to it, the smarter it gets. Except after spending less than one day on Twitter this chat bot became racist and which I think it demonstrates firstly what might be found on Twitter, but also is quite shocking that the AI was able to become racist in less than a day.

And the second thing I came across is that the US government perform some tests on facial recognition systems and it found that even those that were labelled as top performing, they misidentified black people at five to 10 times higher rate than white people and you can begin to imagine what sort of issues this could cause with misidentification of people by facial recognition software.

I'm sure most people who have social media will have experienced ad targeting where they can, you can clearly see that this ad has been targeted to you based on your demographic. And Facebook has actually been sued by a department of the US government on its ad targeting. Those that discriminate by gender, race and religion, as these are protected classes. But it's something that we sort of take for granted in our lives every day.

Google developed an AI hate speech detector, which seems like a good idea in the first place but it turns out to actually be racially biased itself, which is a problem. This is designed to monitor hate speech on social media but the tool turns out to be twice as likely to flag posts by African-Americans as any other race. Just obviously worrying.

And then these last two in the pictures were quite big on social media, so I'm not sure if anyone might have seen them. They caused quite a lot of outrage. The first one on the left with the soap dispenser, this is an employee from the Facebook offices and they posted this video of a soap dispenser that with dispense soap for when a white person's hand was placed into the sensor, but it

did not dispense soap when a black person's hand was placed into the sensor, suggesting that the infrared sensors were not designed to detect dark skin.

And finally, the last example to highlight here is the, possibly the most shocking. The Google Photos image recognition algorithm in 2015 identified black people and classified them as gorillas. This is in 2015 which was very shocking at the time. But actually through some more research: Google hasn't fixed this, they just blocked the image recognition algorithms from classifying anything as a gorilla. Which doesn't actually remove the problem, it just saves them from getting any more bad press I would argue.

So on this point of racial biases, racial biases are thoroughly ingrained in society and they have the potential to be exacerbated in algorithms. This can have very worrying consequences, such as in the criminal justice system. Significant problems here include the lack of unbiased historical data, an unbalanced workforce, and limited user testing. And I would say it is important to note here, and think about what's going on in society at the moment, that it won't be possible to entangle racial biases in AI without society and individuals first becoming anti-racist.

At this point I would again like to flag the equality4success Twitter page as a good resource, starting resource for individuals wanting to become anti-racist. There's lots of good resources there, as can be found all over the internet, as I'm sure most of you will already have come across.

So, what causes AI bias? I've covered most of these already going through the different examples, but I wanted to just put them here together in one place and go through them one by one.

So the first cause is an incomplete or a skewed training set. So, this is when demographic categories are missing and then models fail to scale properly. An example here is the Bride Problem. So this example comes from a website called Image Net, which actually during my research cropped up quite a lot. So Image Net is a website that contains 14 million labelled images but actually 45% of them from the USA which does lead to biased outcomes. So, for example, a photograph like the bride on the left wearing a white dress is labelled as a bride. The North Indian woman wearing a wedding sari or a lehenga is labelled as performance art, rather than a bride.

So this leads us on to the second point which is labels used for training can cause AI bias. So most AI is what's known as supervised machine learning. This means that data is labelled to teach the model how to behave and it's, in most circumstances, humans that create these labels and people, as we've discussed earlier, frequently exhibit biases, both consciously and unconsciously. So this is a second reason for a cause of AI bias.

The third, which we've not touched on too much, but is in the features and the modelling techniques themselves. So, for example, the measurements which are used as inputs for machine learning and the actual model itself can introduce bias. So two examples here.

The first is in text to speech and speech to text technology. So, for example, closed captioning. So these technologies tend to perform poorly for speakers who are women, compared to men. So, when the way in which speech is analysed and modelled is in a way that is more accurate for taller speakers with longer vocal cords and lower pitched voices, which means that this normally, means that it works better for male voices than female voices.

And the second example is in Google Translate. So when you're translating a language that doesn't use gendered pronouns into a language like English which does use gendered pronouns, the default for Google Translate is to use, 'he said' rather than any other pronoun like, 'she said,' or 'they said.'

And this is because the phrase ‘he said,’ occurs more commonly on the internet than any of the others.

So this firstly reflects a faulty algorithm, which just selects the most used pronoun rather than any other method, but it also demonstrates a data based bias here, looking at the frequency of the phrases on the Internet.

So, not to leave it all in doom and gloom: obviously there's lots of ways to address bias in AI and there's different levels here that people with different connections to working in AI can be involved in.

So the first example, obviously, is to ensure diversity in the training samples, as we've seen lots of the biases come from having historical data sets. So that's an important one to address.

The second here is to make sure that the humans who are labelling the samples come from diverse backgrounds. So I did a little bit of research on gender and the tech sector in the UK. And I found that and in 2017 there was a report that showed only 7% of students in computer science were women. And in the workforce, even though women are 47% of the UK workforce, only 17% of the tech sector were women.

So there's obviously a lot of work to go, to do in this area. There are some organizations that are addressing this issue. So two that I've come across that we should definitely support are Black in AI and also the Alan Turing Institute has set up Women in AI society to support these groups of people into tech careers.

The last two are more about the generation of AI. So the first is that you can encourage the machine learning teams to measure the accuracy levels separately for the different demographic categories. This then allows you to identify when one category is being treated unfair, unfavorably.

And there's also something I thought was very interesting called counterfactual fairness that can be applied. This makes, this helps to make the algorithm led decisions fair by making sure that the outcomes are the same when they're in the actual world, but also when they're in what's known as a counterfactual world where an individual belongs to a different demographic group.

And finally, more training data associated with an underrepresented group can be collected and then machine learning techniques in debiasing can be applied that penalize for producing unfairness. And I think these are all important avenues for us to support

So I just wanted to finish on a quote from the Harvard Business Review which puts this into some context, saying that, ‘bias is all of our responsibility. It hurts those discriminated against, of course, but it also hurts everyone by reducing people's ability to participate in the economy and society. It reduces the potential of AI for business and society by encouraging mistrust and producing distorted results. Business and organizational leaders need to ensure that the AI systems they use improve on human decision-making, and that they have a responsibility to encourage progress on research and standards that will reduce bias in AI.”

So I hope that if any of the people listening today currently have any say in factors to do with AI or if in the future you're ever in that position that you might remember some of the things from this presentation to think about when making decisions in this area. I'd like to finish by highlighting some interesting resources which are used to do my research for this presentation. So there's quite a few papers and, there's particularly of interest, there's this report AI and gender: four proposals for

future research by the University of Cambridge. So if it's anything that people are particularly interested in, this might be a local avenue for getting involved.

And then also, again, the link here is to the Harvard Implicit Association test if anyone was interested in finding out a bit more about that.

And with that, I will end the presentation and if anyone has any questions, comments that they would like to make, please post in the chat or unmute yourself.